Optimizing Edge AI: Performance Engineering in Resource-Constrained Environments

Giuliano Casale
Department of Computing
Imperial College London
United Kingdom
g.casale@imperial.ac.uk

ABSTRACT

Recent years have witnessed the growth of Edge AI, a transformative paradigm that integrates neural networks with edge computing, bringing computational intelligence closer to end users. However, this innovation is not without its challenges, especially in environments with limited computing, network, and memory constraints, where resource-hungry AI models often need to be partitioned for distributed execution. This issue becomes even more acute in scenarios where post-deployment updates are infeasible or costly, posing a need to accurately reason about the interplay between resource constraints and Quality-of-Service (QoS) in Edge AI systems, so as to optimally design and operate them.

In this keynote talk, I will focus on these challenges, discussing QoS management and deployment problems arising in Edge AI systems. I will review mechanisms such as early exits and DNN partitioning that are distinctive of this problem space, explaining how they could be accounted for and leveraged in system performance and reliability tuning. I will then illustrate how design decisions and the definition of novel runtime control algorithms can be guided by approaches based on both traditional analytical models and emerging data-driven methods based on machine learning models.

CCS CONCEPTS

• **General and reference → Performance**; • **Computing methodologies → Modeling and simulation**; • **Computer systems organization → Dependable and fault-tolerant systems and networks**.

KEYWORDS

Edge AI, performance, reliability, tuning

ACM Reference Format:
Giuliano Casale. 2024. Optimizing Edge AI: Performance Engineering in Resource-Constrained Environments. In *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24), May 7–11, 2024, London, United Kingdom.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3629526.3649131

ICPE '24, May 7–11, 2024, London, United Kingdom

ACM ISBN 979-8-4007-0444-4/24/05
https://doi.org/10.1145/3629526.3649131

SPEAKER BIO

Giuliano Casale is a Reader in the Department of Computing at Imperial College London. He does research in Quality-of-Service engineering and cloud computing, topics on which he has published more than 150 refereed papers. He has served as program co-chair for several conferences in the area of performance and reliability engineering, such as ACM SIGMETRICS/Performance and IEEE/IFIP DSN. His research work has received best paper awards at ACM SIGMETRICS, IEEE/IFIP DSN and IEEE INFOCOM. During 2019-2023, he has served as ACM SIGMETRICS chair. He serves on the editorial board of ACM TOMPECS and as Editor-in-Chief of Elsevier Performance Evaluation.